Assessing Multiple Myeloma Risk: Differential Expression Analysis and Modeling

Kristen Vossler | Gibson Lab | Georgia Institute of Technology

Readers:

Dr. Greg Gibson

Dr. Chrissy Spencer

Angela Mo, Candidate for Ph.D

Kristen Vossler 4/25/19 10:06 PM
Deleted: Signature

Kristen Vossler 4/22/19 3:20 PM

Deleted: _____
_____ ... [1]

Unknown
Formatted: Font:11 pt

Kristen Vossler 4/22/19 3:20 PM
Deleted: _____ ... [2]

Kristen Vossler 4/25/19 10:06 PM
Deleted:

Kristen Vossler 4/22/19 3:20 PM
Deleted: _____
_

KEY TERMS

Multiple Myeloma: a cancer of the plasma cell component of the blood

Transcriptomics: study of the transcriptome, a measure of the mRNA transcript abundance in a cell as an estimate for gene expression

RNA Sequencing (RNA-Seq): transcriptomics method to determine abundance of each mRNA in a sample via RNA isolation, cDNA synthesis, and alignment to a human reference genome

HTSeq: Python package that takes values from RNA-Seq and calculates the number of reads for each gene, thus determining the level of expression for each gene

Principal Component (PC) Analysis: statistical transformation used to capture major sources of variance in a dataset

Co-expression Matrix: an undirected graph that clusters genes with similar expression values across samples together

Gene Ontology Enrichment: classifies a set of genes by enrichment of biological function

Linear Regression Model: models relationship between a dependent variable and one or more explanatory variables; in this case, expression levels predicting clinical outcomes

Receiver-Operating Curves (ROC): visualization of the predictive capability of a classifier system or model; plots specificity and sensitivity

Sensitivity: proportion of true positives correctly identified by a classifier or model

Specificity: proportion of true negatives correctly identified by a classifier or model

ABSTRACT

Multiple myeloma is an incurable blood cancer affecting more than 20,000 Americans each year[1]. Sub-classification of heterogeneous subsets of the disease under the broad umbrella of multiple myeloma is likely to increase efficacy of treatments. Our study aimed to analyze an open-access transcriptomics dataset to elucidate how gene expression can predict health outcomes in multiple myeloma, principally focusing on overall survival and physical ability. The Multiple Myeloma Research Foundation's CoMMpass dataset utilized RNA sequencing (RNA-Seq) to obtain abundance counts for each of over 20,000 genes expressed in bone marrow. We extracted the 13,000 most abundantly expressed genes at time of diagnosis in 767 patients for whom medication history, time to death, and numerous clinical measures were also available. After normalization and quality control, we performed differential expression analysis as well as one-way ANOVA to correlate expression values with clinical factors. Finally, we constructed a logistic regression model to predict clinical outcome. At a sensitivity of 30% and a specificity of 95% cutoff, we can predict death within 5 years from baseline expression with a precision of 55% for 59 patients, which represents a three-fold higher rate than in the total cohort. This study serves as an important foundation in personalized medicine for multiple myeloma that should be expanded upon by determining if there is a way to increase survival for the high-risk individuals.

INTRODUCTION

Multiple myeloma is the third most common blood cancer in the United States[2]. The disease is characterized by malignancy of plasma cells, white blood cells that produce antibodies, growing without control in the bone marrow. Tumor manifestation leads to severe clinical outcomes such as anemia, bone lesions, renal malfunction, and hypercalcemia[3]. Although multiple myeloma is considered incurable, identification of robust biomarkers of disease progression should support personalized treatments of the disease. Because multiple myeloma is a heterogeneous disease, meaning that the dysregulation of multiple alleles and genes converges on the same clinical phenotype, it has been challenging to discover an effective therapeutic target[4-5].

Traditionally, multiple myeloma is classified either by ploidy, the number of chromosome pairs, or by chromosomal translocations, a phenomenon in which a segment of one chromosome is attached to a new location on the same or a different chromosome[3]. In response to these classifications, multiple treatments have emerged, from steroids to immunomodulators to proteasome inhibitors[5]. Despite this extensive list of treatment options, patients often relapse or fail to respond[6].

Treatments could be significantly improved if they were tailored to the individual genomic profile of differential expression that is driven by key mutations. Transcriptomics, the measure of the number of mRNA transcripts in a cell as a representation of gene expression, provides insight into the cellular pathogenesis of disease. The objective of our study was to use transcriptomics to correlate differential gene expression levels with clinical outcomes in multiple myeloma. We utilize the CoMMpass Multiple Myeloma Research Foundation RNA-Seq database, which consists

Kristen Vossler 4/25/19 9:49 PM
Deleted: Since m

Kristen Vossler 4/25/19 9:50 PM
Deleted: [4]

Kristen Vossler 4/25/19 10:13 PM
Deleted: database which

of pre-computed output files with baseline expression levels for bone marrow samples from 767 multiple myeloma patients at time of diagnosis, with follow-up biopsies for fewer than 100 individuals. Using co-expression matrices and gene ontology enrichment, we sub-classify patients based on similar expression profiles. Furthermore, principal component analysis and one-way ANOVA are used to highlight specific genes correlated with clinical variables and then a linear regression model is used to predict disease outcome. With these tools, this study provides a prediction of prognosis as well as a sub-classification of patients based on gene expression.

Kristen Vossler 4/25/19 9:51 PM
Deleted: use these to construct a linear regression model

LITERATURE REVIEW

Multiple Myeloma

Approximately 20,000 Americans are newly diagnosed with multiple myeloma each year[1]. Worse, most patients who enter remission eventually experience a recurrence, calling into question the overall effectiveness of current treatment methods[6]. To improve treatment efforts, it is necessary to first understand how the disease manifests itself. Plasma cells are the final differentiated state of B-lymphocytes, the white blood cells that produce antibodies for the immune system. Plasma cells can become cancerous after control over regulation of the cell cycle is lost due to somatic mutations, and continued abnormal cell division results in malignancy. When the tumor forms in the bone marrow, the abnormal cells produce an antibody called monoclonal (M) protein[3]. This first stage is known as monoclonal gammopathy of undetermined significance (MGUS) when no organ or skeletal damage has yet occurred[7]. Because MGUS is asymptomatic, patients are not screened for the M protein unless a family history is present[5]. 20% of MGUS patients progress to multiple myeloma[8]. The patients at highest risk for deterioration have primary genetic factors, such as hyperdiploidy or chromosomal translocations[3]. Hyperdiploidy is due to an abnormal mitotic cycle, resulting in chromosomal instability[9]. Translocations occur spontaneously from double strand DNA breaks that lead to rearrangements of the immunoglobulin heavy chain genes (IgH). The reshuffle sometimes indirectly results in the up-regulation of enhancers controlling oncogenes, promoting cell growth[5]. MGUS can develop directly into multiple myeloma or undergo an intermediate phase known as smoldering multiple myeloma (SMM)[3]. The disease is finally classified as multiple myeloma when numerous tumors have formed in the bone marrow. The tumors continue

Kristen Vossler 4/25/19 9:53 PM
Deleted: or hyperdiploidy

to expand, applying force onto the skeleton and surrounding organ systems that results in lesions and pain[7].

Beyond symptoms, there are two testing criteria systems for multiple myeloma. The first benchmarks are known as the CRAB criteria, which stands for increased calcium, renal failure, anemia, and bone lesions. Bone lesions can be detected through imaging such as an MRI, PET scan, or CT[7]. Extensive blood work is also performed to determine chemical composition. Another standard for diagnosis is the International Staging System (ISS) that measures the severity of the disease on a scale of 1-3 with 3 being the most critical. Levels of Beta 2 Microglobulin and albumin distinguish the stages[10]. If a patient meets both CRAB and ISS criteria, he or she will be diagnosed with multiple myeloma.

To improve the criteria measures, studies are being performed to characterize biomarkers and identify causal factors. Multiple myeloma is a heterogeneous disease, meaning that while the phenotype appears the same across patients, the genetic network varies significantly among individuals[3]. As with most cancers, the heterogeneity makes it difficult to determine causal drivers of multiple myeloma[11]. However, there are two clear markers of poor prognosis in multiple myeloma: t(4;14) and del(17p13). The more substantial marker is del(17p13) as it disrupts p53, a key regulator of cell cycle checkpoints[3]. Specific significantly involved genes include *KRAS, NRAS, TP53, BRAF,* and *FAM46C*[12]. Crucial pathways include NF-kB activation, apoptosis, the cell cycle, *Wnt* signaling, and histone modifications, similar to other forms of cancer[13]. However, these factors have not yet proven to be sufficient to significantly improve treatments, likely because they are characteristic of most cancers and do not help with staging. An

Kristen Vossler 4/25/19 9:53 PM
Formatted: Indent: First line: 0.5"

Kristen Vossler 4/25/19 9:54 PM
Deleted: Other s

improved classification system and prediction model would improve treatment and diagnostic methods.

Transcriptomics

The transcriptome offers a potential novel approach to stratification of risk[14-15]. By generating sub-categories of multiple myeloma based on differential gene expression, treatment might be tailored to individuals on the basis of inappropriately active pathways. In addition, sub-categories may differ in time to death or survival rates, allowing patients and families to adjust their planning. The initial focus of research has widely been on genomic mutations[11] through whole exome sequencing.

The two major techniques for measuring the transcriptome are microarrays and RNA sequencing (RNA-Seq)[16]. Microarrays determine transcript abundance through hybridization of the mRNA transcripts to complementary probes deposited on a silicon wafer, usually obtained from the companies Illumina or Affmetrix[16]. RNA-Seq more directly measures transcript abundance by sequencing short fragments of cDNA. The sequences are then aligned to a reference genome, and the number of reads associated with each gene is computed. These are called read-counts, or simply counts[17]. Although it is a more recent technology and was initially expensive, costs have dropped to less than $500 per sample. Most researchers now prefer RNA-Seq because it is able to evaluate splicing variants and non-coding RNAs, and is not biased by the content of the array[17].

After retrieving the counts, differential gene expression is performed using a statistical program such as the SAS product JMP Genomics[18] or one of the open source R packages (edgeR[19] or DESeq2[20]) to normalize, model, and identify the differentially expressed genes[21]. By correlating gene expression values with clinical factors,

researchers can predict outcomes of a disease such as overall survival or treatment efficacy[22]. Gene expression can also stratify patients into groups with similar expression patterns, helping to break down the umbrella phenotype into more specific subtypes[22]. Once genes of interest are identified, gene ontology enrichment permits classification of the molecular function so that researchers can reach a deeper understanding of the mechanisms at play[23].

MATERIALS AND METHODS

The purpose of our research was to determine if there are significant differences in gene expression associated with overall survival and physical ability in multiple myeloma patients. Utilizing these differences, we constructed a sub-classification of the disease as well as linear regression models to predict risk.

Gene Expression Data for Multiple Myeloma Patients

We obtained expression data from the Multiple Myeloma Research Foundation's CoMMpass dataset. The counts were generated from RNA-Seq data using the HTSeq Python package and were taken directly from the CoMMpass website. We extracted samples from the bone marrow at time of diagnosis as a proxy for baseline expression levels. We included the 10,000 most abundant genes for 767 de-identified patients. We logarithmically transformed the data (log2) in order to measure the differences in expression on a scale that naturally allows the interpretation of 2-fold differences, while not being dominated by a small number of high-abundance genes. Table 1 shows the race, age, and gender of the study participants.

Kristen Vossler 4/25/19 10:14 PM
Deleted:

Kristen Vossler 4/25/19 9:55 PM
Deleted: Multiple Myeloma Research Foundation

Race	(n)	**Age**	(n)
White	*501*	20-35	*3*
Black	*99*	35-50	*76*
Asian	*12*	50-65	*335*
Other	*35*	65-80	*277*
Missing	*120*	80-95	*54*
Gender	(n)	Missing	*22*
Male	*387*		
Female	*260*		
Missing	*120*		

Table 1. CoMMpass Dataset Demographics

Kristen Vossler 4/25/19 10:14 PM
Deleted: .

Normalization and Principal Component Analysis

The data was processed through the Basic Expression Workflow in JMP-Genomics, with mean-centered normalization. Principal component (PC) analysis was used to capture the major components of variance in the dataset. The first 8 PC's each captured more than 2% of the variance each; subsequent PCs captured less and were therefore not used in estimates. We then correlated the first 8 PCs with clinical factors to determine whether covariates of interest (survival, ECOG performance, and ISS) significantly contributed to transcriptional variance.

Differential Gene Expression and Gene Ontology Enrichment

We used one-way ANOVA to identify significantly differentially expressed genes for survival and ECOG performance at the threshold of $p<0.001$. Two-way hierarchical clustering of these genes was performed using Ward's method in order to sub-classify

patients. We then conducted gene ontology enrichment using the ToppGene Suite[24] to annotate the molecular and biological functions of the significant genes. We subsequently correlated the first two PCs of overall survival and ECOG performance to determine if the two were associated with one another. Finally, we used a paired t-test to determine whether expression levels change over time in 40 patients who were included in the dataset at two different times during the progression of their multiple myeloma.

Kristen Vossler 4/25/19 9:57 PM
Deleted: for whom RNA-Seq data at two time points was available.

Predictive Models

In order to identify a baseline predictive score which performs better than random assignment, we performed logistic regression on demographic measures, including age, race, ethnicity, and gender. We then constructed a logistic regression model using the first 10 PC's to predict survival within 5 years. The optimal level of specificity and sensitivity to obtain the highest negative and positive predictive values was acquired by examination of the Receiver-Operating Curve (ROC).

RESULTS

Normalization and Quality Control Processes

Mean-centered normalization was effective in reducing the among-sample variability as seen by comparing the top and bottom panels in Figure 1. This effectively ensures that the total number of counts per sample is constant. Before normalization, log2 counts ranged between 0 and 20; after normalization they are centered on 0.

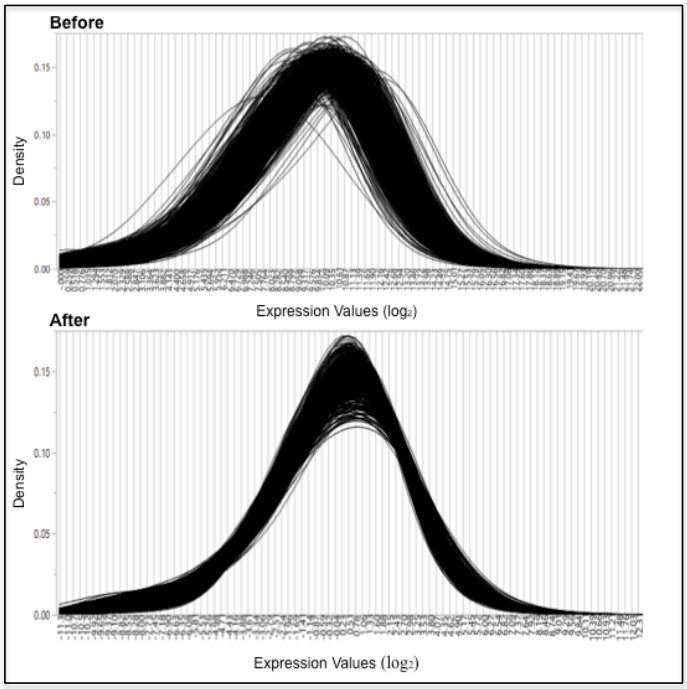

Figure 1. Mean-Centered Normalization

The panel on top is the distribution of counts without normalization processes. On the bottom, the distribution has been tightened and centered on 0 to control for individual differences in the dataset.

Principal Variance Component Analysis

Principal variance component analysis identifies the strongest sets of covariance in a dataset, and once obtained these can be assessed for their contribution to variance of factors of interest. The first 8 PC collectively explained 33% of the total variance, and since subsequent PCs each explain <2%, so they were excluded from further consideration. The weighted average proportion of variance explained across principal components demonstrated that gender, overall survival (OS), age, race, ECOG Performance (in five categories, see below, Figure 5):, and ISS (International Staging System score) each explain less than 5% of the variance explained by the 8 PCs, and collectively just 11.6%, corresponding to 3.8% of the total variance (Figure 2). The remaining residual variance is due to technical, environmental, among individual, and unidentified biological sources. The proportion of variance relative to the clinical scales (ISS, ECOG Performance, and OS) is shown in Figure 3.

Kristen Vossler 4/25/19 10:00 PM
Deleted: p

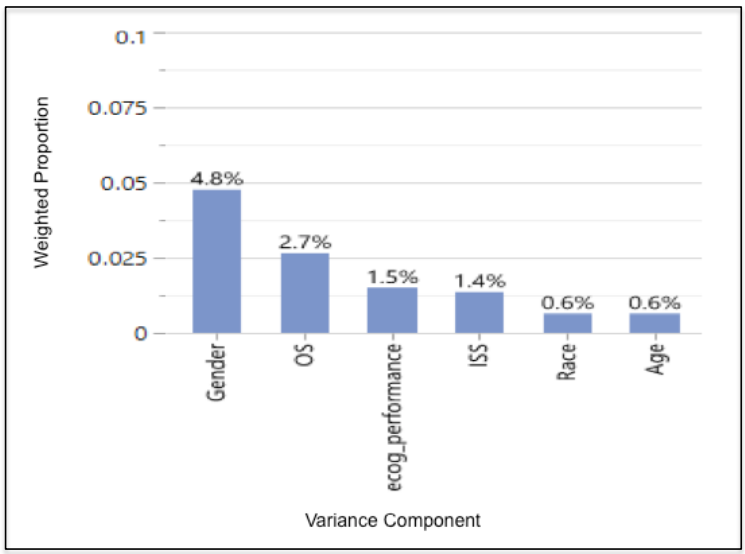

Figure 2. Weighted Proportion Across Variance Components

The figure shows how much each clinical variable contributes to variance in the dataset, with gender being the most significant source of variance (4.8%).

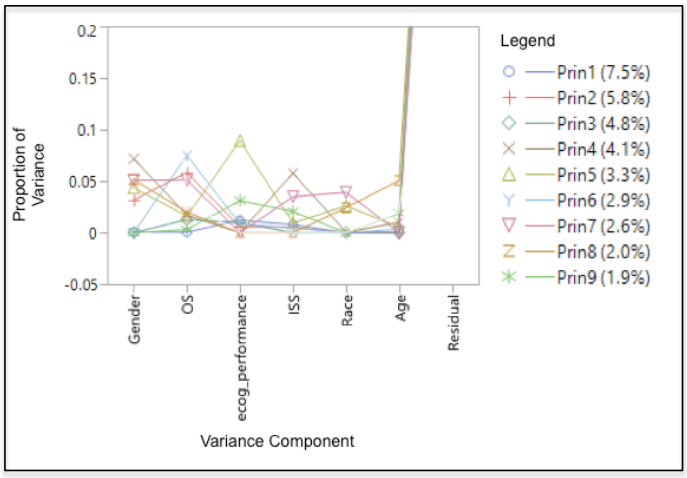

Figure 3. Variance Proportion by Principal Component

The figure estimates which PCs are most related to each factor, as well as shows how much variance each individual PC captures.

When we correlated the first 8 PCs with clinical factors, we found that PCs 2, 6, and 7 were significantly correlated with overall survival (OS: p = 0.0001, 0.0001, and <0.0001 respectively; Figure 4). PC5 was reflective of ECOG Performance, a measure of physical ability (p < 0.0001; Figure 5), and the ISS disease severity score was associated with PC4 and PC7 (p = 0.0007, <0.0001 respectively).

Kristen Vossler 4/25/19 10:15 PM
Formatted: Indent: First line: 0"

Kristen Vossler 4/25/19 10:15 PM
Deleted: . ... [3]

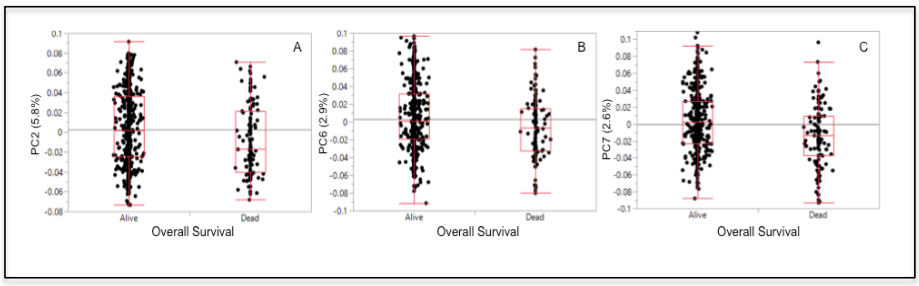

Figure 4. Significant Correlations with Overall Survival

We performed a t-test to determine significant differences in overall survival for the first 8 PCs. We show here the three significant PCs, all of which demonstrated a higher score was more favorable for survival. A) PC2 (p = 0.0001) B) PC6 (p = 0.0001) C) PC7 (p < 0.0001)

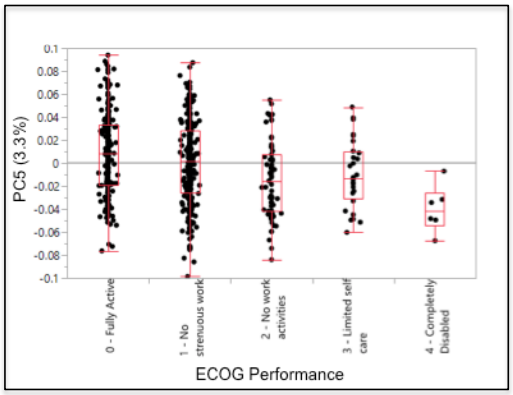

Figure 5. Significant Correlation with ECOG Performance

We performed an ANOVA test to determine significant differences in ECOG Performance for the first 8 PCs, finding PC5 to be the only significant component ($p < 0.0001$). A higher value for PC5 is associated with better physical activity. The five categories are 0 - Fully active, 1 – Strenuous work not allowed, 2 – No work activity allowed, 3 – Limited self-care needed, and 4 – Completely disabled.

Differential Gene Expression Analysis

When we performed two-way hierarchical clustering on the top 13,000 genes (Figure 6), the patients fell into nine distinct clusters. We also found three groups of differentially expressed genes (columns) among the clusters and analyzed their gene classifications through gene ontology (Tables 2, 3, and 4). Patient cluster 1 had the best prognostic signature, as most patients were alive, fully active, and in Stage 1 of ISS. Contrastingly, patient cluster 9 appears to have the worst signature with most patients deceased within 5 years, not able to perform work activities, and in Stage 2 or 3.

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	Cell adhesion molecule binding	6.254E-19	88	208
	Signaling receptor binding	5.475E-18	391	1601
	Actin binding	1.389E-16	135	412
Biological Process				
	Biological adhesion	3.284E-68	522	1542
	Cell adhesion	1.151E-67	518	1530
	Immune response	1.480E-57	505	1572
Cellular Component				
	Intrinsic component of plasma membrane	2.632E-36	476	1714
	Cell surface	7.820E-36	289	873
	Integral component of plasma membrane	4.484E-34	456	1651
Pathway				
	Cytokine signaling in immune system	7.070E-25	243	763
	Cell adhesion molecules (CAMs)	1.565E-19	72	145
	Innate immune system	3.620E-19	346	1312

Table 3. Gene Group 1 (3843 Genes) for Top 13,000 Genes

	Name	p-value	Genes in input	Genes from annotation
Molecular Function				
	RNA binding	1.650E-87	961	1632
	Structural constituent of ribosome	2.794E-38	170	216
	Transferase activity	5.644E-19	146	225
Biological Process				
	RNA processing	4.782E-79	600	913
	ncRNA metabolic process	3.783E-77	415	563
	ncRNA processing	7.983E-67	310	399
Cellular Component				
	Mitochondrion	9.139E-107	1055	1769
	Mitochondrial part	1.237E-89	650	987
	Ribonucleoprotein complex	5.604E-68	492	745
Pathway				
	Gene expression	1.722E-126	1103	1844
	Cell cycle	8.489E-49	392	624
	Organelle biogenesis and maintenance	3.762E-44	243	341

Table 4. Gene Group 2 (7574 Genes) for Top 13,000 Genes

Molecular Function	Name	p-value	Genes from input	Genes in annotation
	RNA binding	1.576E-45	331	1632
	Ubiquitin-like protein transferase activity	1.120E-18	103	441
	Enzyme binding	3.132E-18	293	1929
Biological Process				
	RNA processing	1.611E-33	204	913
	Protein modification	1.866E-32	218	1026
	RNA splicing	3.399E-28	114	402
Cellular Component				
	Nucleoplasm part	2.490E-35	178	732
	Transferase complex	1.452E-24	155	722
	Catalytic complex	2.668E-21	195	1071
Pathway				
	Gene expression	1.474E-25	300	1844
	Processing of capped intron containing pre-mRNA	3.187E-16	67	248
	mRNA splicing	2.691E-14	55	188

Table 5. Gene Group 3 (1966 Genes) for Top 13,000 Genes

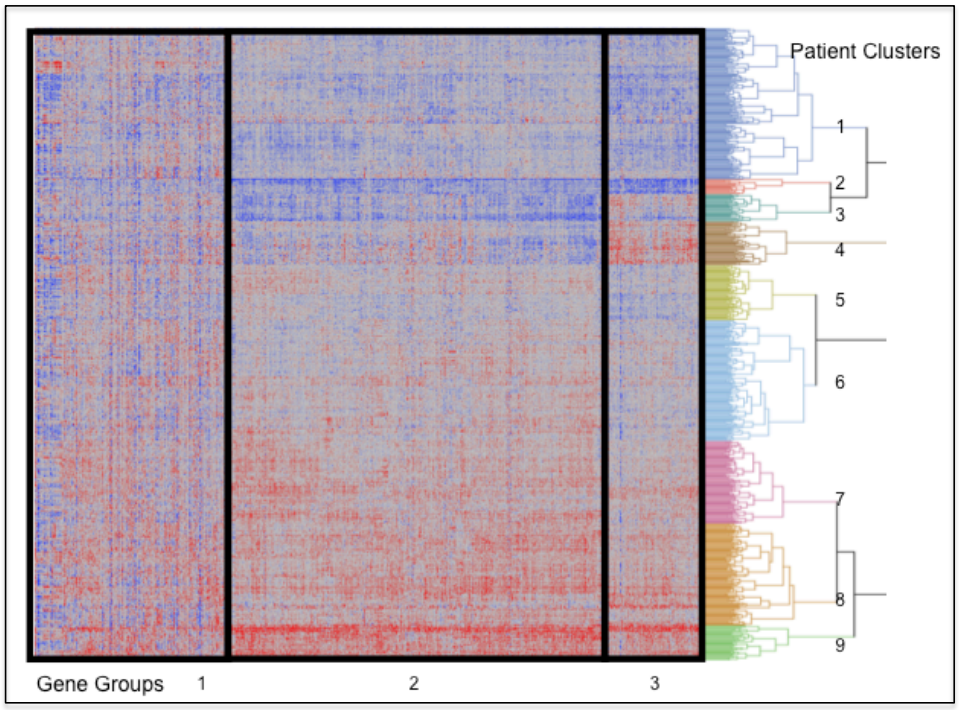

Figure 6. Two-way Hierarchical Clustering Using Ward's Method

The patients fell into nine clusters of individuals (rows), denoted by numbers 1-9. The differentially expressed genes separated into three major groups (columns), marked 1-3. The interior red represents high levels of expression while the blue indicates low levels of expression.

Overall Survival

We then used one-way ANOVA to determine the significant differences in expression related to overall survival. The volcano plot displays how the magnitude of differential expression is related to the significance of difference (Figure 7).

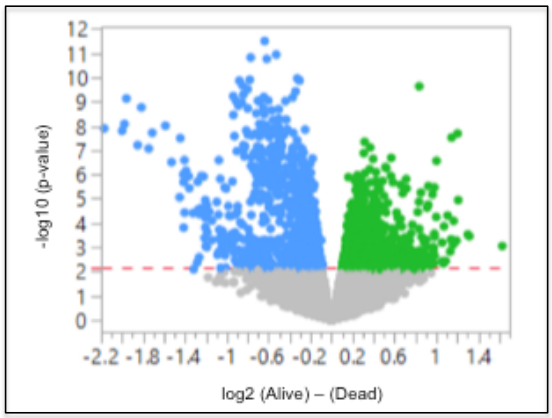

Figure 7. Volcano Plot of Overall Survival

The volcano plot displays the extent of differentially expressed genes between patients who survive and those who are deceased. The genes more highly expressed in deceased patients are found on the left, colored blue. The genes more highly expressed in living patients are found on the right, colored green. Nominally significantly different genes lie above the red dashed line (NLP = 2.1).

When we clustered the genes highly associated with survival and genes highly associated with death (Figure 8), we found 8 patient clusters and 6 gene groups (Tables 6-11). Patient clusters 4 and 7 appear to have the worst outcomes associated with their expression profiles, such as death and limited self-care. Patient clusters 2, 3, and 8 contain mostly patients who are alive and fully active.

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	Cytoskeletal protein binding	2.631E-8	34	886
	Protein binding involved in heterotypic cell-cell adhesion	6.005E-6	4	10
	Calmodulin binding	9.353E-6	12	190
Biological Process				
	Movement of cell or subcellular component	2.551E-7	52	1882
	Cardiac muscle cell-cardiac muscle cell adhesion	1.041E-6	4	7
	Actin filament based process	1.666E-6	26	685
Cellular Component				
	Cell junction	3.508E-7	38	1200
	Cell surface	3.947E-6	29	873
	Neuron part	4.271E-6	42	1545

Table 6. Gene Group 1 (281 Genes) for Overall Survival Cluster

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	RNA binding	1.598E-5	49	1632
	ADP binding	4.685E-5	6	41
Biological Process				
	Cell cycle process	1.206E-10	56	1385
	Mitotic cell cycle	6.286E-9	43	1016
	Mitotic cell cycle process	1.497E-8	40	931
Cellular Component				
	Chromosome	2.534E-6	35	941
	Centromeric region	8.335E-6	13	185
	Condensed chromosome	4.436E-5	9	105
Pathway				
	Cell cycle	2.20E-9	32	624
	Gene expression	2.795E-9	61	1844
	Mitotic cell cycle	1.235E-7	26	517

Table 7. Gene Group 2 (322 Genes) for Overall Survival Cluster

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	DNA helicase activity	4.339E-11	9	65
	Nucleoside-triphosphatase activity	5.583E-11	22	785
	ATP binding	1.411E-10	29	1475
Biological Process				
	Cell cycle process	5.999E-58	70	1485
	Mitotic cell cycle process	6.383E-57	62	931
	Cell cycle	1.613E-56	74	1766
Cellular Component				
	Chromosome	8.036E-36	47	941
	Chromosomal part	2.209E-30	41	839
	Chromosomal region	2.176E-26	28	344
Pathway				
	Cell cycle	9.564E-46	49	624
	Mitotic cell cycle	1.042E-38	42	517
	Mitotic prometaphase	9.035E-25	20	111

Table 8. Gene Group 3 (97 Genes) for Overall Survival Cluster

	Name	p-value	Genes from input	Genes from annotation
Cellular Component				
	Vacuolar part	2.585E-7	47	700
	Vacuolar membrane	1.849E-6	40	593
	Golgi membrane	6.411E-6	44	716
Pathway				
	Lysosome	1.499E-6	15	123
	Trafficking and processing of endosomal TLR	2.586E05	5	14

Table 9. Gene Group 4 (687 Genes) for Overall Survival Cluster

Biological Process	Name	p-value	Genes from input	Genes from annotation
	Glycosylceramide metabolic process	8.706E-7	5	14
	Glucosylceramide metabolic process	1.200E-6	4	7
	Response to hydroperoxide	3.561E-6	5	18

Table 10. Gene Group 5 (322 Genes) for Overall Survival Cluster

	Name	p-value	Genes from input	Genes from annotation
Molecular Function				
	Phospholipid binding	1.521E-5	6	385
	Phosphatidylcholine-sterol-O-acyltransferase activator activity	3.013E-5	2	6
	Cargo receptor activity	1.446E-4	3	71
Biological Process				
	Inflammatory response	4.023E-12	13	711
	Defense response	4.253E-12	17	1651
	Endocytosis	1.534E-8	10	665
Cellular Component				
	Extracellular space	2.941E-9	14	1449
	Plasma lipoprotein particle	3.282E-7	4	40
	Protein-lipid complex	4.011E-7	4	42
Pathway				
	Interleukin-4 and 13 signaling	4.294E-5	4	114
	Altered lipoprotein metabolic	6.226E-5	2	7
	Staphylococcus aureus infection	1.249E-4	3	56

Table 11. Gene Group 6 (115 Genes) for Overall Survival Cluster

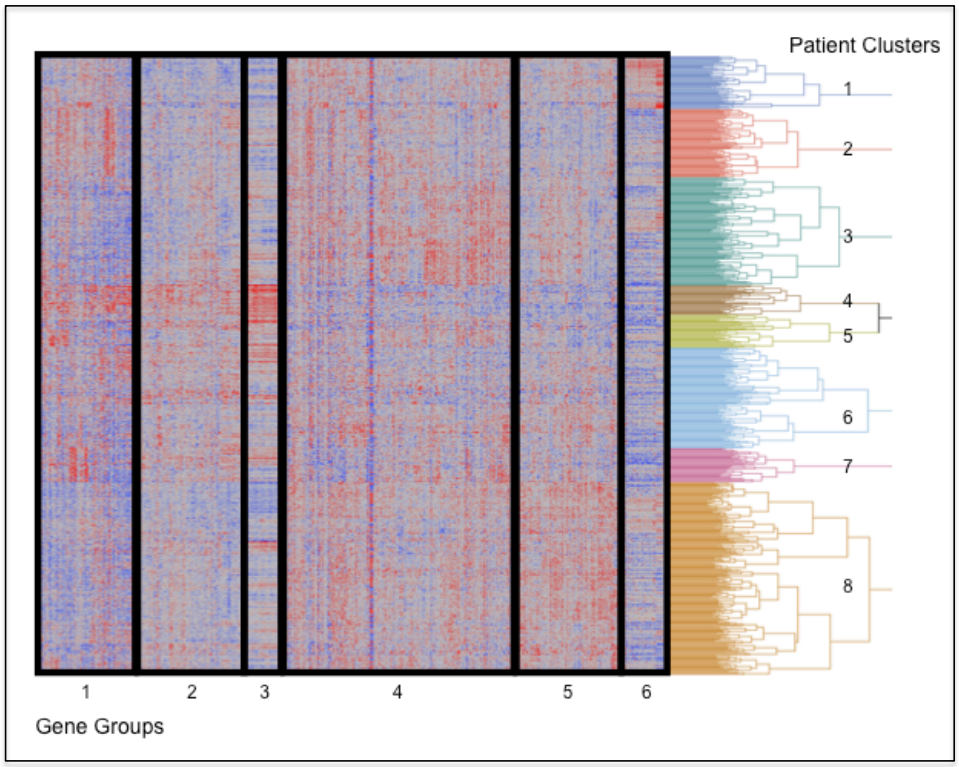

Figure 8. Two-way Hierarchical Clustering of Overall Survival

The patients fell into eight clusters of individuals (rows), denoted by numbers 1-8. The differentially expressed genes separated into six major groups (columns), marked 1-6. The interior red indicates high levels of expression; the blue represents low levels of expression.

ECOG Performance

We then performed one-way ANOVA to evaluate differences in expression related to ECOG Performance status. ECOG Performance has five levels: Fully Active (34.5%), No strenuous work allowed (48.0%), No work allowed (11.6%), Limited self-care (4.9%), and Completely disabled (1.0%). We generated volcano plots of each

pairwise comparison, and found that the most significant differential expression distinguished Limited self-care and Fully active individuals (Figure 9).

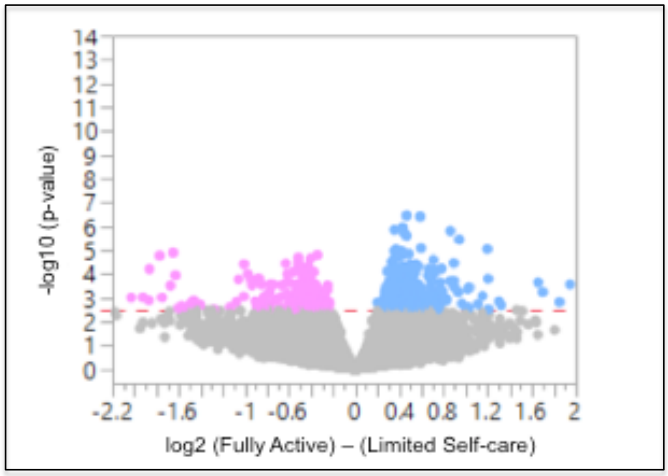

Figure 9. Volcano Plot of Fully Active and Limited Self-Care Genes

The volcano plot identifies significantly differentially expressed genes between patients who are Fully active and those who have Limited self-care abilities. The genes more highly expressed in patients with Limited self-care are found on the left peak, colored pink. The genes more highly expressed in Fully active patients are found on the right peak, colored blue. Significant genes lie above the red dashed line (NLP = 2.5).

When we clustered genes highly associated with being fully active and those associated with limited self-care (Figure 10), we found eight patient clusters and five gene groups (Tables 12-14). Gene group 2 (66 genes) and gene group 3 (63 genes) were not significant enough to garner hits for our gene ontology categories. Patient clusters 1, 2, 4, and 5 appear to have the best signatures as the patients are mostly alive, fully active or conducting no strenuous work, and are in Stage 1 or Stage 2. Patient cluster 7 is

Kristen Vossler 4/22/19 3:23 PM
Deleted: s

associated with the worst prognosis, with most patients being deceased, or not able to perform any work activities, and Stage 3.

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	Ubiquitin-protein transferase activity	3.716E-9	14	414
	Ubiquitin-like protein transferase activity	7.651E-6	14	441
	Zinc ion binding	3.231E-4	21	1200
Biological Process				
	Protein modification by small protein conjugation or removal	2.088E-6	23	102
	Chromosome organization	7.385E-6	24	118
	Protein modification by small protein conjugation	1.248E-5	20	90
Cellular Component				
	Transferase complex	1.596E-6	19	722
	Catalytic complex	4.105E-6	23	1077
	Nucleoplasm part	7.843E-6	18	732
Pathway				
	YAP-1 and WWTR1 gene expression	4.027E-5	4	29
	Signaling by NOTCH	1.119E-4	6	113

Table 12. Gene Group 1 (166 Genes) for ECOG Performance

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	RNA binding	5.315E-14	46	1632
	snRNP binding	1.920E-5	3	7
	NADH dehydrogenase activity	3.860E-5	5	46
Biological Process				
	Ribonucleoprotein complex biogenesis	1.770E-18	30	468
	DNA metabolic process	9.584E-13	34	992
	Ribosome biogenesis	2.487E-12	20	322
Cellular Component				
	Mitochondrial inner membrane	4.99E-14	26	507
	Organelle inner membrane	8.045E-14	27	564
	Mitochondrial envelope	1.863E-13	30	736
Pathway				
	Huntington's disease	1.991E-10	15	193
	Respiratory electron transport, ATP synthesis by chemiosmosis	1.415E-9	12	126
	Oxidative phosphorylation	2.644E-9	12	133

Table 13. Gene Group 4 (166 genes) for ECOG Performance

	Name	p-value	Genes from input	Genes in annotation
Molecular Function				
	Leptomycin B binding	1.180E-3	1	1
	Glycine C-acetyltransferase activity	1.180E-3	1	1
	Neurotrophin TRKC receptor binding	1.180E-3	1	1
Cellular Component				
	Anchored component of external side of plasma membrane	3.470E-4	2	23
	External side of plasma membrane	4.759E-4	4	310
	Intrinsic component of external side of plasma membrane	5.544E-4	2	29

Table 14. Gene Group 5 (29 Genes) for ECOG Performance

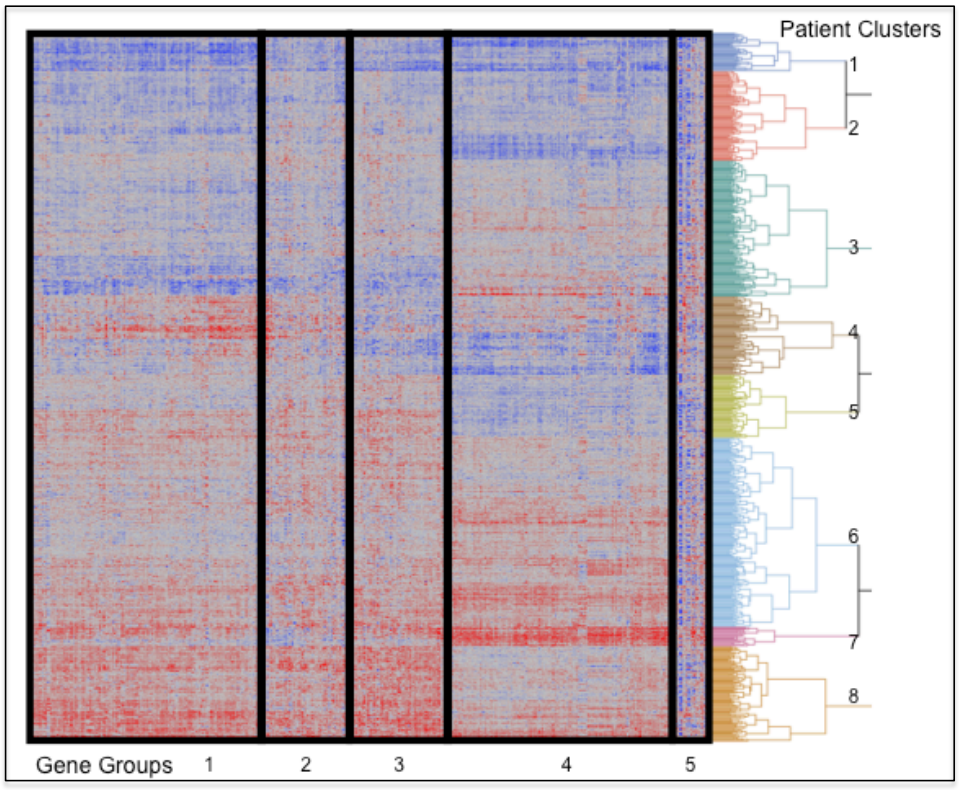

Figure 10. Two-way Hierarchical Clustering of ECOG Performance

The patients fell into eight clusters of individuals (rows), denoted by numbers 1-8. The differentially expressed genes separated into six major groups (columns), marked 1-6. The interior red represents high levels of expression while the blue indicates a low level of expression.

Correlation of Overall Survival and ECOG Performance

We found that PC2 of overall survival and PC2 of ECOG Performance were linearly correlated ($R^2 = 0.26$; Figure 11); however, other pairings were not significant, with $R^2 < 0.1$. This indicates that Overall Survival and Physical Activity are only

partially correlated, and with the exception of the second two PCs, appear to associate with different aspects of plasma cell gene expression.

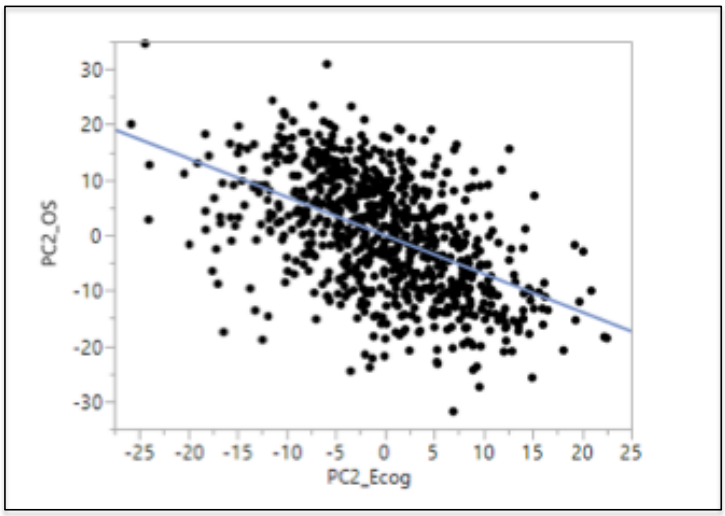

Figure 11. Bivariate Fit of Overall Survival and ECOG Performance

PC2 of overall survival and PC2 of ECOG Performance are linearly correlated (R^2 = 0.26). A higher level of one results in lower score of the other.

Expression over Time

We found that expression levels do not significantly alter over time; in fact, they stay quite stagnant as seen through a paired t-test (Figure 12) for 40 patients with two measurs, as well as a t-test of the average expression score (Figure 13) between all of the baseline samples and the second timepoint for the 40 patients. This implies that time of sampling is not likely to substantially influence prediction of survival, and that gene expression profiles do not become worse as disease progresses.

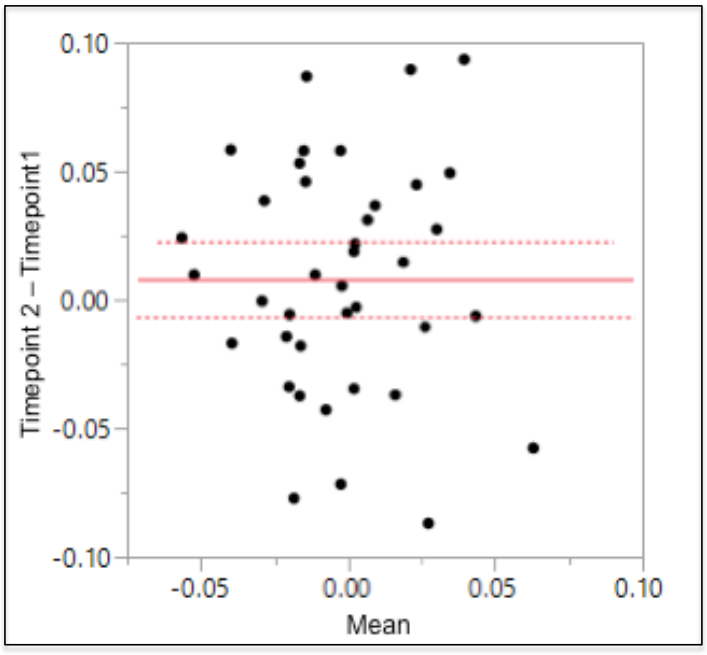

Figure 12. Paired t-test

The paired t-test allowed us to compare indivual's expression levels at baseline and at a second time point. We found that expression scores stay similar to baseline. The average expression level is shown by the thick red line.

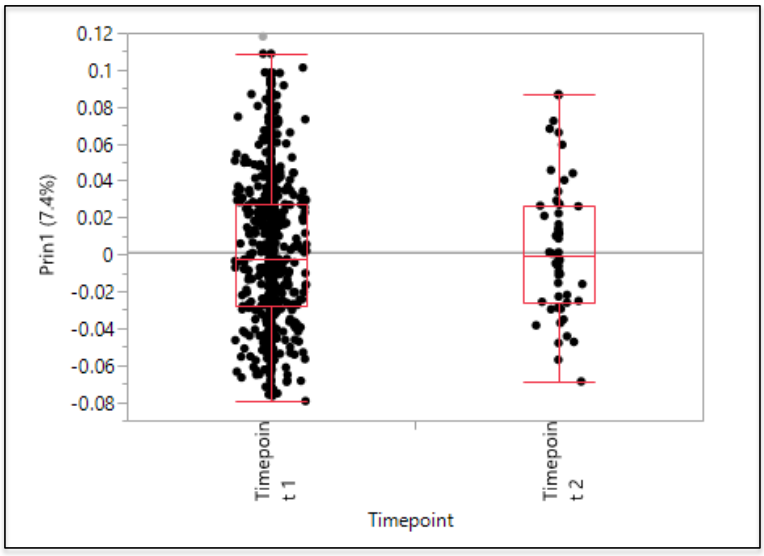

Figure 13. Comparison of Average Expression Over Time

We found no significant difference between the average expression levels of timepoint 1 and timepoint 2.

Predictive Accuracy of Models

We constructed a logistic regression model using the first 10 PCs in order to form a baseline predictor of 5-year survival (Table 15). We compared three different points on the ROC curve, determining our best cut-off to focus on the most at-risk individuals (Figure 14). The accuracy at this point was 84% with 95% specificity but only 30% sensitivity. 54% of the 59 patients called positive would die within 5 years, compared to 17% of the total set of participants for whom data was available (107/646); and 87% of those called negative would survive, compared with 80% from the available data. The positive prediction represents a three-fold increase in the likelihood of early death, information which may be useful for families and patients planning the final years.

Kristen Vossler 4/25/19 10:20 PM
Deleted: . ... [4]

30% Sensitivity and 90% Specificity	Cases	Controls	Predicted Value
Called Positive	*32*	*27*	*PPV = 0.55*
Called Negative	*75*	*512*	*NPV = 0.87*

70% Sensitivity and 65% Specificity	Cases	Controls	Predicted Value
Called Positive	*75*	*189*	*PPV = 0.29*
Called Negative	*32*	*350*	*NPV = 0.92*

90% Sensitivity and 20% Specificity	Cases	Controls	Predicted Value
Called Positive	*97*	*431*	*PPV = 0.18*
Called Negative	*11*	*108*	*NPV = 0.91*

Table 15. Accuracy Table for Logistic Regression Classification Using First 10 PCs

The table data is derived from the ROC Curve found in Figure 14 taking the 3 cutoff points at various sensitivity and specificity levels.

Kristen Vossler 4/22/19 3:25 PM
Formatted ... [5]
Kristen Vossler 4/22/19 3:28 PM
Formatted Table ... [6]
Kristen Vossler 4/22/19 3:25 PM
Formatted ... [7]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [8]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [9]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [10]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [11]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [12]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [13]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [14]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [15]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [16]
Kristen Vossler 4/22/19 3:28 PM
Formatted Table ... [17]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [18]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [19]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [20]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [21]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [22]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [23]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [24]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [25]
Kristen Vossler 4/22/19 3:28 PM
Formatted ... [26]
Kristen Vossler 4/22/19 3:28 PM
Formatted Table ... [27]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [28]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [29]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [30]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [31]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [32]
Kristen Vossler 4/22/19 3:29 PM
Formatted ... [33]
Kristen Vossler 4/22/19 3:29 PM
... [34]
Kristen Vossler 4/25/19 10:17 PM
... [35]

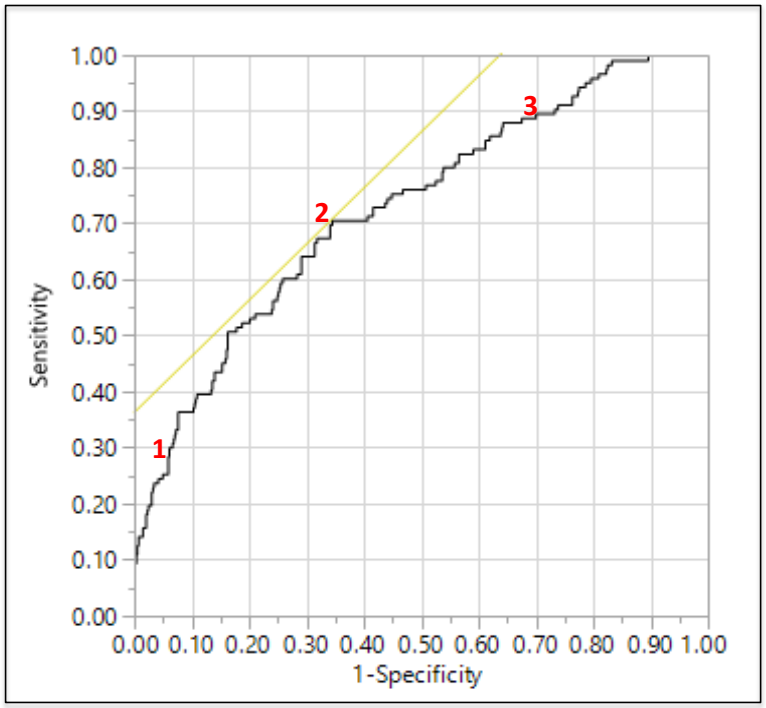

Figure 14. ROC Curve Using First 10 PCs

Death within 5 years was selected as the positive level when constructing the ROC curve. The curve identifies the three points we compared for positive predictive value, negative predictive value, and accuracy. Point 1 was the optimal classifier for predicting death within 5 years (PPV = 0.55; NPV = 0.87; Accuracy = 0.84), followed by Point 2 (PPV = 0.29; NPV = 0.92; Accuracy = 0.66), and then finally Point 3 (PPV = 0.18; NPV = 0.91; Accuracy = 0.32).

DISCUSSION

Our analyses revealed significant differential expression among multiple myeloma patients in bone marrow samples at baseline diagnosis. Through principal component analysis, we found 1.5% of the variance was from gender differences in the samples. Gender was strongly associated with PC10 ($p < 0.0001$), and it was shown that males have worse genetic signatures than females. We also found that overall survival explained 0.9% of the variance, while physical ability, measured by ECOG Performance, was associated with 0.5% of the variance. After correlating the principal components with the clinical and demographic measures, PC7 became of major interest, as it was significantly associated with overall survival and ISS ($p=0.0001$; $p=0.0002$). Higher scores of PC7 were more correlated with survival and better outcomes. However, the majority of the variance was residual, and PC1 only captured 7.5% of the total variance. This is not unusual in gene expression profiling experiments since technical and among individual variation is prevalent.

When we clustered the data, we found that high expression levels of gene groups 1, 2, and 3 are correlated with worse outcomes in the patients. However, it more represents how the patients separate at a physiological level. Gene group 1 is related to cell adhesion. Previous literature has demonstrated the link between increased adhesion and multiple myeloma progression. By adhering to the bone marrow microenvironment, the tumor receives physical support and growth factors – allowing further growth and spread of the cancer[25]. Researchers are looking into disrupting the adhesion molecules as an additional form of therapy through antibodies and signaling inhibition[26]. Gene group 2 is related to protein translation, while gene group 3 is associated with post-transcriptional

modification to RNA. An increase in changes to RNA transcripts can result in translational reprogramming, working in cancer's favor by taking advantage of these regulatory mechanisms. For example, alternative splicing can result in non-functional proteins as well as changes that can result in spread of the cancer[27]. There could also be an affinity for microRNAs, which block the synthesis of apoptotic proteins, allowing for the progression of tumor formation[28]. Furthermore, an increase in translation means more proteins are being translated – whether pro-apoptotic or cancerous. Overall, these three groups represent the basic physiology clusters among multiple myeloma patients' plasma cell expression profiles.

Kristen Vossler 4/25/19 10:04 PM
Deleted: however

We then clustered the significantly differentially expressed genes specifically in relation to survival; we found that high expression of gene groups 1, 2, and 3 are correlated with negative outcomes. Gene group 1 is associated with calmodulin binding and actin filaments, gene group 2 is correlated with the cell cycle, and gene group 3 is related to ATP binding and chromosomal instability. Calmodulins mediates Ca^{2+}, and as calmodulin increases in a cell, the cell cycle progresses. Yokokura, S. and colleagues found that by inhibiting calmodulins, they could induce apoptosis in cancerous cells – a potentially useful treatment for multiple myeloma[29]. Chromosomal instability and the cell cycle are common hallmarks of cancer, giving weight to our findings. We also found that low expression of gene groups 4, 5, and 6 were associated with worse outcome. Gene group 4 is related to the endomembrane system and gene group 5 is associated with metabolism. The endomembrane system packages proteins to be used in the membrane or outside of the cell. Down-regulation of this system could decrease intracellular communication. One study showed that bone marrow mesenchymal cells regularly

Kristen Vossler 4/25/19 10:05 PM
Deleted: One study

communicate with plasma cells through extracellular vesicles, supporting and maintaining plasma cell survival[30]. Disruption of this communication could lead to stress on the cell, resulting in worse outcomes. Contrastingly, other studies have demonstrated that multiple myeloma can be diagnosed through evidence of increased amounts of extracellular vesicles in the cell, contradicting our argument. Higher levels of extracellular vesicles correlate with increased tumor-growth promoting signals[31]. However, while elevated levels of extracellular vesicles may be an effective diagnostic marker, it may not be indicative of survival or physical ability. As for gene group 5, by down-regulating pathways related to metabolism, a tumor cell is able to control what metabolites are accumulated in the cell. While previous studies show an increase in metabolites related to glucose and glutamine, this surge could be relative to down-regulation of other metabolites[32]. For example, osteoprotegerin, an inhibitor of osteoclast activator RANK, has significantly lower expression in multiple myeloma patients[33]. This decrease results in increased osteoclast functions, breaking down the surrounding bone.

Kristen Vossler 4/25/19 10:18 PM
Deleted:

Subsequently, we clustered genes associated with being ECOG Performance. We found that up-regulation of gene group 4 is associated with limited self-care or disability. Gene group 4 is correlated with cellular respiration. Mitochondrial biogenesis and oxidative phosphorylation has been shown to be elevated in multiple myeloma patients with worse outcomes in a study that also examined gene expression profiles[34]. The increased expression may result in mitochondrial chromosome instability, as well as induce cellular stress.

We then correlated PCs related to overall survival and ECOG Performance. While PC1 of overall survival and PC1 of ECOG Performance were not significantly correlated,

as well as the rest of the pair wise comparison, PC2 of overall survival and PC2 of ECOG Performance were correlated somewhat. This indicates that physical ability and survival are related, but do not drive one or the other.

When we compared baseline expression levels of 40 patients over time, we found there was no significant change, meaning our evaluations at baseline are just as clinically useful as 6 to 12 months following diagnosis, as the expression measure do not vary substantially. We can define sub classifications of patients and treat them specifically for their cluster without worrying about whether the expression will alter significantly.

When we constructed our 5-year survival prediction model, we found that the cutoff at 30%% sensitivity and 95% specificity was optimal for accuracy and positive predictive value, the two measures most clinically relevant. This cutoff only includes 59 of the patients, namely less than 10%, but with a positive predictive value of 55%, we will be able to predict correctly half of the people who will die from baseline expression levels – three times greater than the 17% given in the dataset. With a negative predictive value of 87%, we will be able to predict most of the people who will survive, although this is not a marked improvement on the 80% who are survivors at 5 years in any case. More inclusive cutoffs have lower precision and accuracy.

Overall, our results indicate a few major conclusions. First, there is a significant difference in gene expression in plasma cells of multiple myeloma patients. These differences are mainly defined by adhesion, translation, and post-transcriptional modification. Second, higher expression of calmodulin binding, chromosomal instability, and cellular respiration in bone marrow plasma cells of multiple myeloma patients all associate with increased risk of early death. Third, low expression of pathways related to

the endomembrane system and metabolism also leads to elevated risk. Last, our classifier of 5-year survival using the first 10 PCs should be used to predict prognosis from diagnosis. Those with higher risk should receive more extreme and diligent treatment procedures, while those with lowest risk may have a more positive prognosis and be able to plan accordingly.

Kristen Vossler 4/27/19 6:14 PM
Deleted: m

Kristen Vossler 4/25/19 10:18 PM
Formatted: Indent: First line: 0"

REFERENCES

1. Rajkumar S.V., and S. Kumar. 2016. Multiple myeloma: Diagnosis and treatment. *Mayo Clin Proc* **91**: 101-119.
2. Foundation MMR. 2018. What is multiple myeloma; https://themmrf.org/multiplemyeloma/what-is-multiplemyeloma/
3. Fonseca, R, et al. 2009. International Myeloma Working Group: Molecular classification of multiple myeloma. *Leukemia* **23**: 2210-2221.
4. De Mel, S., Lim, S.H., Tung, M.L., and W.J. Chng. 2014. Implications of heterogeneity in multiple myeloma. *Biomed Res Int* **2014**: 232546.
5. Morgan, G.J., Walker, B.A., and F.E. Davies. 2012. The genetic architecture of multiple myeloma. *PLoS Comput Biol* **13**: e1005457.
6. Naymagon, L, and M. Abdul-Hay. Novel agents in the treatment of multiple myeloma: a review about the future. *J Hematol Oncol* **9:** 52.
7. Kyle, R.A., and S.V. Rajkumar. 2009. Criteria for diagnosis, staging, risk stratification, and response assessment of multiple myeloma. *Leukemia* **23**: 3-9.
8. Kumar, S.K., Rajkumar, V., Kyle, R.A., van Duin, M., Sonneveld, P., Mateos, M.V., Gay, F., and K.C. Anderson. 2017. Multiple myeloma. *Nat Rev Dis Primers* **3**: 17046.
9. Carroll, W.L. 2013. Safety in numbers: hyperdiploidy and prognosis. *Blood* **121**: 2374-2376.
10. Fonseca, R., Monge, J., and M. A. Dimipoulos. 2014. Staging and prognostication of multiple myeloma. *Expert Rev Hematol* **7**: 21-31.

11. Uhlen, M., et al. 2017. A pathology atlas of the human cancer transcriptome. *Science* 357.

12. Bolli, N., et al. 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**: 2997.

13. Walker, B.A., et al. 2010. A compendium of myeloma associated chromosomal copy number abnormalities and their prognostic value. *Blood* **116**: e56-65.

14. Wrighton, K.H. 2017. Genetic testing: the diagnostic power of RNA-seq. *Nat Rev Genet* **18**: 392-393.

15. Gibson, G. Powell, J.E., and U.M. Marigorta. 2015. Expression quantitative trait locus analysis for translational medicine. *Genome Med* **7**: 60.

16. Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and T. Shafee. 2017. Transcriptomics technologies. *PLoS Comput Biol* **13**: e1005457.

17. Soekojo, C.Y., de Mel, S., Ooj, M., Yan, B., and W.J. Chang. 2018. Potential clinical application of genomics in multiple myeloma. *Int J Mol Sci* 19.

18. JMP®, Version 8, SAS Institute Inc., Cary, NC, 1989-2019.

19. Robinson, M.D., McCarthy, D.J., and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140.

20. Love, M.I., Huber, W., and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* **15**: 550.

21. Shaheen, S., Fawaz, F., and S. Shah. 2018. Differential expression and pathway analysis in drug-resistant triple-negative breast cancer cell lines using RNA-Seq analysis. *Int J Mol Sci* **19**(6): 1810.

22. Seckinger, A., Bahr-Ivacevic, T., Benes, V., and D. Hose. 2018. RNA-Sequencing from low-input material in multiple myeloma for application in clinical routine. *Methods Mol Biol* **1792**: 97-115.

23. Chen, J., Bardes, E.E., Aronow, B.J., and A.G. Jegga. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* **37**(1): 305-311.

24. Hunter, M.C., Pozhitkov, A.E., and P.A. Noble. 2017. Accurate predictions of post-mortem interval using linear regression analyses of gene meter expression data. *Forensic Sci Int* **275**: 90-101.

25. Cook, G., Dumbar, M., and I. M. Franklin. 1997. The role of adhesion molecules in multiple myeloma. *Acta Haematol* **97**(1-2): 81-89.

26. Nair, R., Gebhard, A. W., Emmons, M. F., and L. A. Hazlehurst. 2012. Emerging strategies for targeting cell adhesion in multiple myeloma. *Advances in Pharmacology* **65**: 143-189.

27. Jewer, M., Findlay, S.D., and L.M. Postovit. 2012. Post-transcriptional regulation in cancer progression. *J Cell Commun Signal* **6**(4): 233-248.

28. Oliveto, S., Mancino, M., Manfrini, N., and S. Biffo. 2017. Role of microRNAs in translation regulation and cancer. *World J Biol Chem* **8**(1): 45-56.

29. Yokokura, S., Yurimoto, S., Matsuoka, A., Imataki, O., Dobashi, H., Bandoh, S., and T. Matsunaga. 2014. Calmodulin antagonists induce cell cycle arrest and apoptosis in vitro and inhibit tumor growth in vivo in human multiple myeloma. *BMC Cancer* **14**: 882.

30. Nguyen, D. C., Lewis, H. C., Joyner, C., Warren, V., Xiao, H., Kissick, H. T., Wu, R., Galipeau, J., and F. Lee. 2018. Extracellular vesicles from bone marrow derived mesenchymal stromal cells support ex vivo survival of human antibody secreting cells. *J Extracell Vesicles* **7**(1): 1463778.

31. Wang, J., Faict, S., Maes, K., De Bruyne, E., Van Valckenborgh, E., Schots, R., Vanderkerken, K., and E. Menu. 2016. Extracellular vesicle cross-talk in the bone marrow microenvironment: implications in multiple myeloma. *Oncotarget* **7**(25): 38927-38945.

32. Arfani, C. E., Veirman, K., Maes, K., Bruyne, E., and E. Menu. 2018. Metabolic features of multiple myeloma. *Int J Mol Sci* **19**(4): 1200.

33. Seidel, C., Hjertner, O., Abildgaard, N., Heickendorff, L., Hjorth, M., Westin, J., Nielsen, J.L., Hjorth-Hansen, H., Waage, A., Sundan, A., and M. Borset. 2001. Serum osteoprotegerin levels are reduced in patients with multiple myeloma with lytic bone disease. *Blood* **98**: 2269-2271.

34. Zhan, X., Yu, W., Franqui-Machin, R., Bates, M. L., Nadiminti, K., Cao, H., Amendt, B. A., Jethava, Y., Frech, I., Zhan, F., and G. Tricot. 2017. Alteration of mitochondrial biogenesis promotes disease progression in multiple myeloma. *Oncotarget* **8**(67): 111213-111224.